PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-067094
(43)Date of publication of application : 16.03.2001

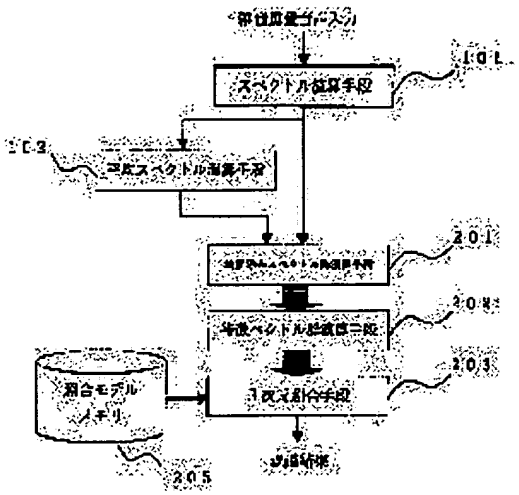
(51)Int.CI. G10L 15/20
G10L 21/02
G10L 15/14

(21)Application number : 11-242856
(22)Date of filing : 30.08.1999

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : NARITA TOMOHIRO

(54) VOICE RECOGNIZING DEVICE AND ITS METHOD

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a voice recognizing device and its method capable of reducing deterioration in recognition performance due to a change in distance between an input terminal of a voice signal and a noise source, and due to variations in environmental noise.
SOLUTION: This voice recognizing device is equipped with a spectrum computing means 101 for obtaining a noise-superimposed voice spectrum time series, an average spectrum computing means 102 for obtaining a noise spectrum by estimating a spectrum of superimposed noise from non-vocal zones, a noiseremoved spectrum group computing means 201 for obtaining noise-removed vocal-spectrum time series of a plurality of kinds by changing a scaling factor relative to the noise spectrum, a characteristic vector group computing means 202 for converting the noise-removed vocal spectrum time series of two or more kinds into characteristic vector time series of two or more kinds, a collation model memory 205 for memorizing a noiseless voice pattern and a model representing transition of the kinds of the characteristic vectors, and a three-dimensional collation means 203 for collating the noiseless voice pattern with the model representing the transition of the kinds of the characteristic vectors in a three-dimensional space made up of three axes, time, state, and characteristic vector.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

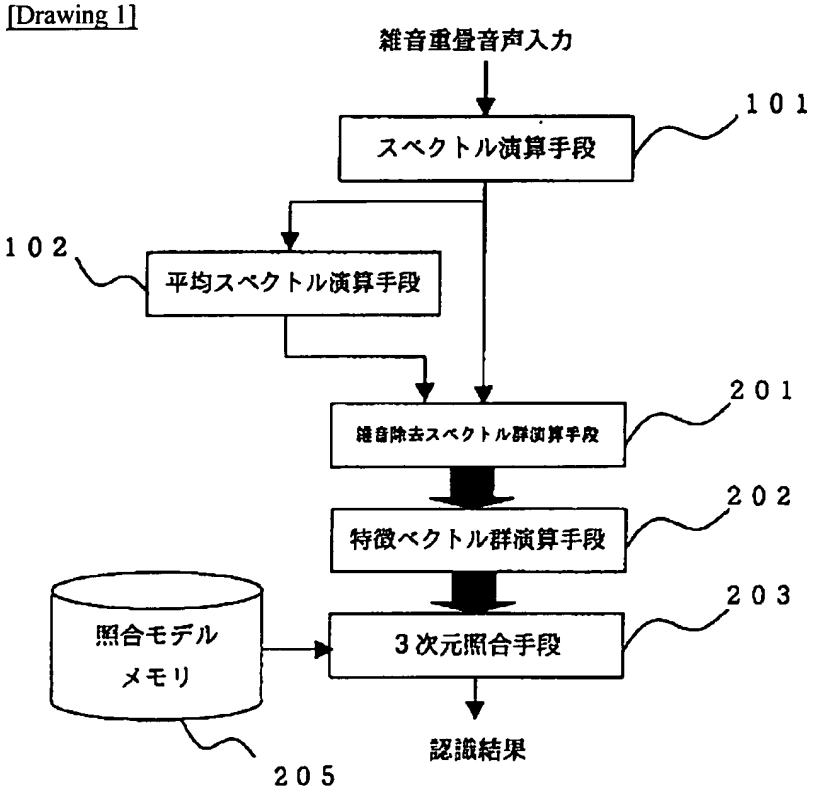
[Date of extinction of right]

* NOTICES *

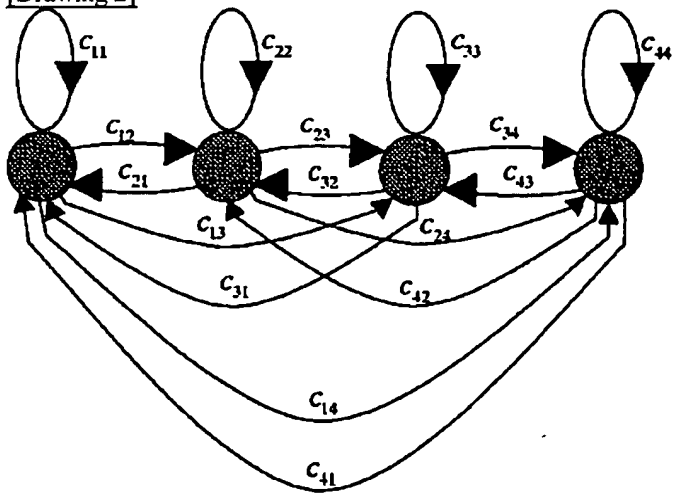
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

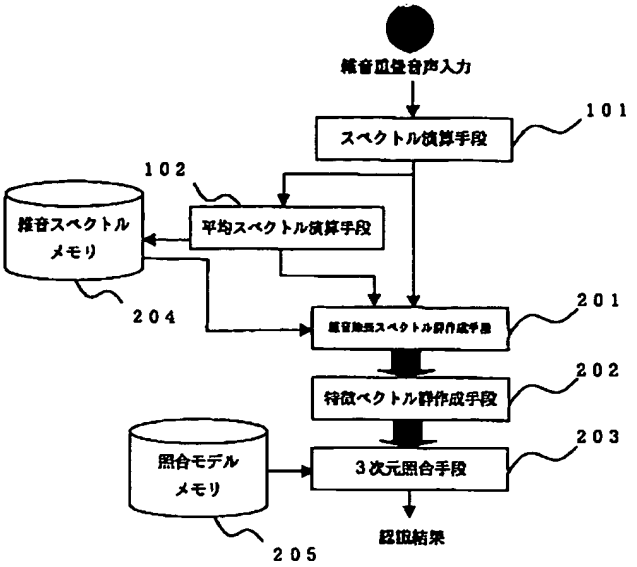
[Drawing 1]



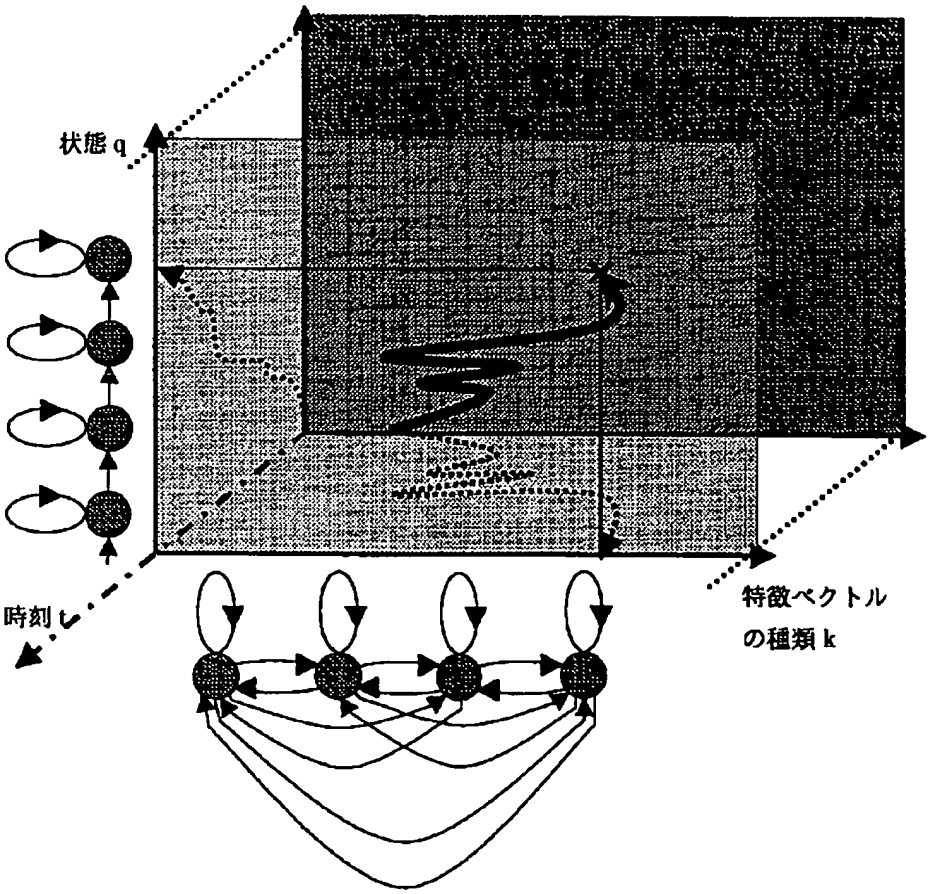
[Drawing 2]



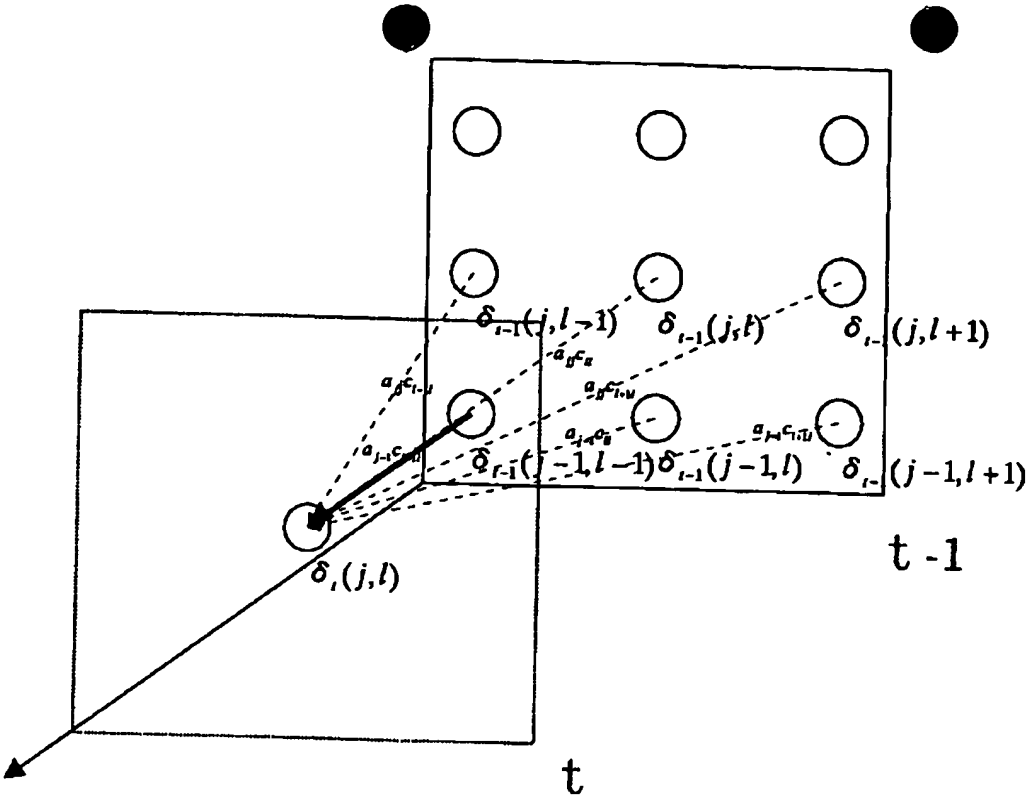
[Drawing 6]



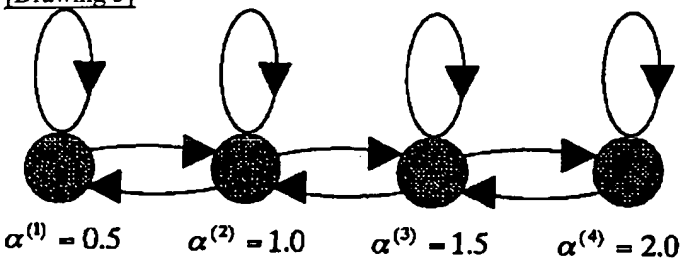
[Drawing 3]



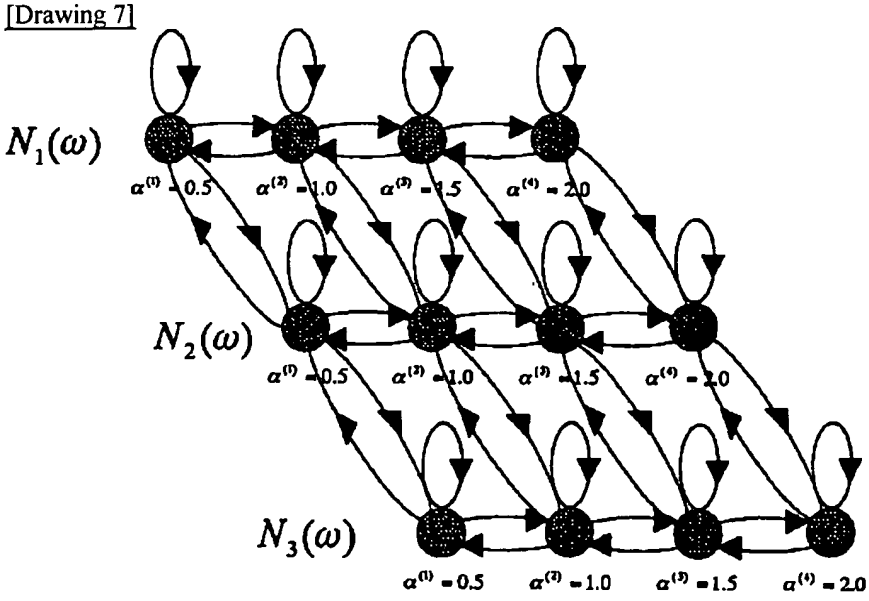
[Drawing 4]



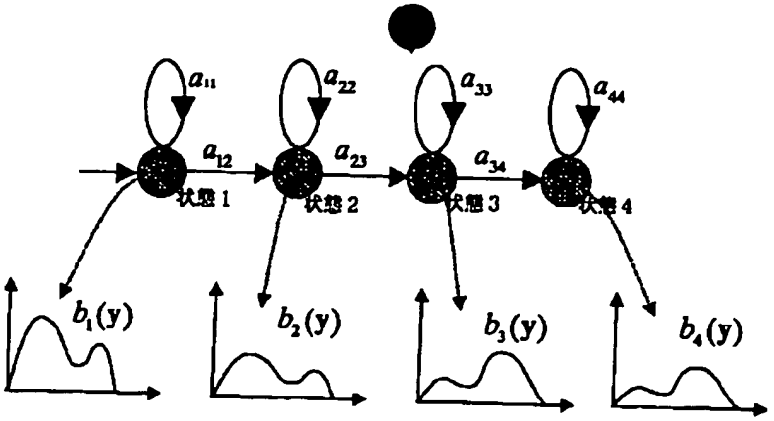
[Drawing 5]



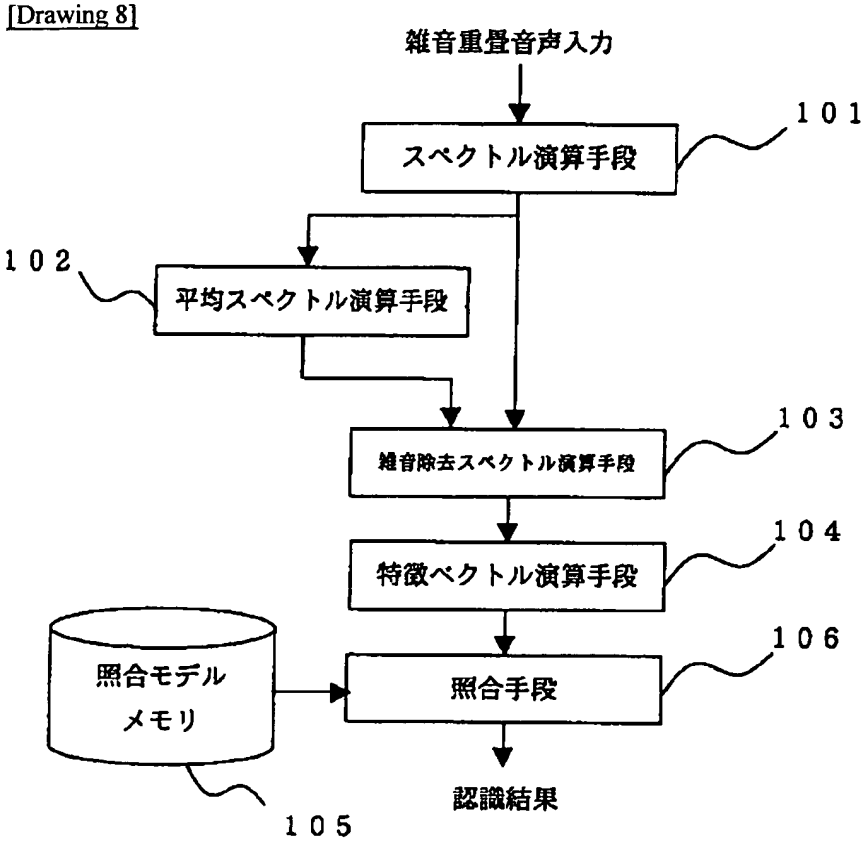
[Drawing 7]



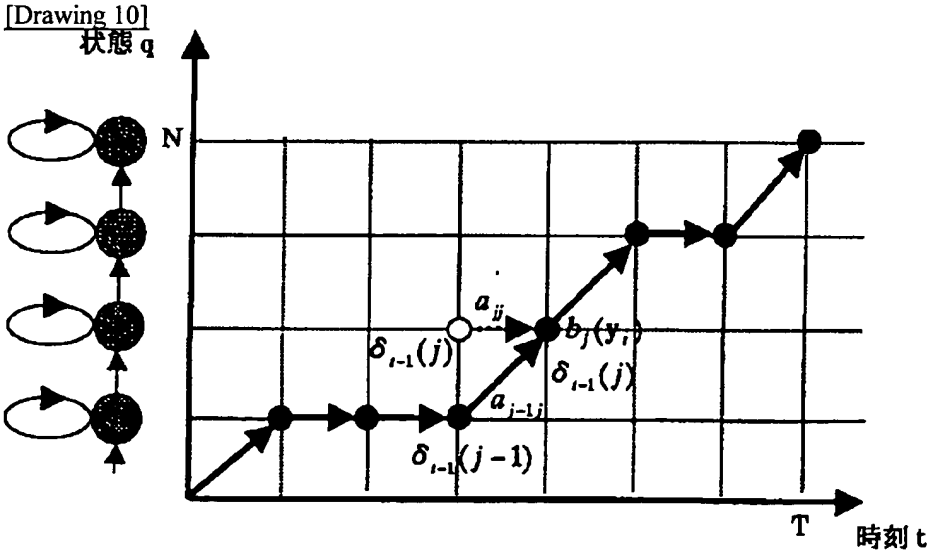
[Drawing 9]



[Drawing 8]



[Drawing 10]



[Translation done.]

* NOTICES *

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is the block diagram showing the composition for explaining the voice recognition unit and method concerning the gestalt 1 of implementation of this invention.
[Drawing 2] It is explanatory drawing of the elgoticHMM model which explains the voice recognition unit and method concerning the gestalt 1 of implementation of this invention, and expresses changes of the kind of feature vector.
[Drawing 3] The voice recognition unit and method concerning the gestalt 1 of implementation of this invention are explained, and it is the state transition of the voice pattern for collating Left-to-right It is explanatory drawing showing the situation of a 3-dimensional Viterbi search in case a HMM model expresses and an elgotic HMM model expresses changes of the kind of feature vector.
[Drawing 4] It is explanatory drawing which extracted the range of the time t-1 in drawing 3 - t.
[Drawing 5] It is explanatory drawing of the HMM model which enabled the changes only of between the kinds of feature vector which explains the voice recognition unit and method concerning the gestalt 1 of implementation of this invention, and adjoins.
[Drawing 6] It is the block diagram showing the composition for explaining the voice recognition unit and method concerning the gestalt 2 of implementation of this invention.
[Drawing 7] It is explanatory drawing of the HMM model which enabled the changes only of between the kinds of feature vector which explains the voice recognition unit and method concerning the gestalt 2 of implementation of this invention, and adjoins.
[Drawing 8] It is the block diagram showing the composition of the voice recognition unit of the conventional example.
[Drawing 9] It is explanatory drawing which expresses the state transition of the voice pattern for collating of the conventional example with the HMM model of Left-to-right which restrictions attached to the state transition.
[Drawing 10] It is explanatory drawing showing the situation of a Viterbi search in case the HMM model of Left-to-right expresses the state transition of the voice pattern for collating.
[Description of Notations]
101 A spectrum operation means, 102 An average spectrum operation means, 201 A normal-mode-rejection spectrum group operation means, 202 A feature-vector group operation means, 203 A 3-dimensional collating means, 204 Noise spectrum memory, 205 Collating model memory.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[The technical field to which invention belongs] This invention relates to the voice recognition unit and method for the voice which it was uttered under noise environment and noise superimposed.
[0002]
[Description of the Prior Art] The background noise is overlapped on the voice uttered under noise environment, and the rate of speech recognition deteriorates. As easy and effective technique for removing this superposition noise, spectrum subtraction technique is used widely. Here, the conventional voice recognition unit using the spectrum subtraction technique indicated by reference "acoustical engineering lecture 7 edited by Acoustical Society of Japan revised voice" (Kazuo Nakada, Corona Publishing Co., Ltd., p.130-131) is explained as the example.
[0003] Drawing 8 is the block diagram showing the composition of the conventional voice recognition unit. The spectrum operation means which 101 performs a analysis of a spectrum to noise superposition voice input, and carries out extract operation of the noise superposition voice spectrum time series in drawing 8 , An average spectrum operation means for 102 to average the spectrum of the non-voice section and to output as a noise spectrum, A normal-mode-rejection spectrum operation means for 103 to subtract a noise spectrum from noise superposition voice spectrum time series, and to output normal-mode-rejection spectrum time series, A feature-vector operation means by which 104 asks for feature-vector time series from normal-mode-rejection spectrum time series, the collating model memory 105 remembers the noise-less voice pattern for collating to be, and 106 receive feature-vector time series. It is a collating means to output the recognition result which performs collating processing with the noise-less voice pattern which the collating model memory 105 memorizes, and gives the maximum likelihood.
[0004] Hereafter, operation of the conventional voice recognition unit is explained. With the spectrum operation means 101, to noise superposition voice input, a power spectrum is calculated by the Fourier transform for every fixed time, and it outputs as time series of a noise superposition voice spectrum. Moreover, with the average spectrum operation means 102, the noise superposition voice spectrum for several frames extracted from the pause section in front of the non-voice section in noise superposition voice spectrum time series, for example, the voice section, or under voice phonation is averaged for every frequency, and it outputs as a noise spectrum. With the normal-mode-rejection spectrum operation means 103, a noise spectrum is subtracted from each noise superposition voice spectrum of the time series of a noise superposition voice spectrum.
[0005] When here shows the relation between the power S in the frequency omega of a normal-mode-rejection voice spectrum (omega), the power X in the frequency omega of a noise superposition voice spectrum (omega), and the power N in the frequency omega of a presumed noise spectrum (omega), it is as a formula (1).
[0006]
[Equation 1]

$$S(\omega) = \max\{X(\omega) - \alpha N(\omega), 0\} \qquad (1)$$

[0007] In addition, alpha is the parameter called sub TORAKUTO coefficient, it expresses the grade which removes a noise component, and usually, it adjusts it so that recognition precision may be made into the maximum. Moreover, max{} is a function which returns the element of the greatest value in the element in a parenthesis.
[0008] The feature-vector operation means 104 is changed into the vector which expresses the acoustical feature in speech recognition, such as an LPC (Linear Predictive Coding) cepstrum, from the normal-mode-rejection voice spectrum time series which the normal-mode-rejection spectrum operation means 103 outputs.
[0009] The collating means 106 performs collating with the noise-less voice pattern which the collating model memory 105 memorizes to the feature-vector time series which the feature-vector operation means 104 outputs, and outputs the recognition candidate who gives a maximum likelihood as a recognition result. Here, the operation method of a maximum likelihood using the Viterbi search in the voice recognition unit using the hidden Markov model (it is called Following HMM) indicated by reference "the foundation (below) of speech recognition" (Lawrence Rabiner, Biing-Hwang Juang collaboration, NTT advance technology incorporated company, p.125-128) is explained as an example of a collating means.
[0010] That is, the Viterbi search which finds one optimum-state sequence q= (q1, q2, ..., qT) which becomes the likelihood

maximum to feature-vector time series Y= (y1, y2, ..., yT) to time 1-T consists of the following four steps.
[0011] STEP1 (initialization)
[0012]
[Equation 2]
$$\delta_1(i) = \pi_i b_i(\mathbf{y}_1) \qquad 1 \le i \le N \qquad (2)$$

[0013]
[Equation 3]
$$\psi_1(i) = 0 \qquad 1 \le i \le N \qquad (3)$$

[0014] STEP2 (repeat)
[0015]
[Equation 4]
$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(\mathbf{y}_t), \quad 2 \le t \le T, \quad 1 \le j \le N \qquad (4)$$

[0016]
[Equation 5]
$$\psi_t(j) = \arg\max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \quad 2 \le t \le T, \quad 1 \le j \le N \qquad (5)$$

[0017] STEP3 (end)
[0018]
[Equation 6]
$$P^* = \max_{1 \le i \le N} [\delta_T(i)] \qquad (6)$$

[0019]
[Equation 7]
$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)] \qquad (7)$$

[0020] STEP4 (backtracking)
[0021]
[Equation 8]
$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \cdots, 1 \qquad (8)$$

[0022] Here, deltat (i) is a maximum likelihood in the time t on the path of one, and is expressed with the following formulas.
[0023]
[Equation 9]
$$\delta_t(i) = \max_{q_1, q_2, \cdots q_{t-1}} P[q_1 q_2 \cdots q_{t-1}, q_t = i, \mathbf{y}_1 \mathbf{y}_2 \cdots \mathbf{y}_t \mid \lambda] \qquad (9)$$

[0024] Formula (2) In - (8), psit (j) is an array which memorizes the argument of the path which makes a formula (9) the maximum in each time t and each state j. Moreover, the output probability of the feature vector [in / State i / aij, and / in bi (yt)] yt, the probability that pii exists in State i by the initial state, and lambda express the voice model for collating, and are learned from the voice data uttered under the environment which does not have noise, respectively. / State / i] / the transition probability to State j
[0025] In a common voice recognition unit, the HMM model of Left-to-right which restrictions attached to the state transition as shown in drawing 9 expresses the state transition of the voice pattern for collating. In addition, bi (y) is the output probability of feature-vector y in State i.
[0026] The situation of a Viterbi search in case the HMM model of Left-to-right expresses the state transition of the voice pattern for collating is shown in drawing 10 . maximum-likelihood deltat- [in / time t-1 and State j / in maximum-likelihood deltat / in / Time t and State j / in drawing 10] (j)] -- calculating by choosing from 1 (j) and maximum-likelihood deltat-1 (j-1) in time t-1 and a state j-1 a path which becomes the likelihood maximum is shown
[0027] By the above operation, it considers that the average spectrum of the noise section of the non-voice section is overlapped on the spectrum time series of the noise superposition sound signal inputted, after removing a noise component on a power spectrum, collating processing with a noise-less collating model is performed, and a recognition result is obtained.
[0028]
[The technical problem which invention tends to solve] Since the bottom voice recognition unit of noise using the

conventional spectrum subtraction method is constituted as mentioned above, when the difference of the noise spectrum superimposed on the average spectrum of the noises in front of utterance etc. and the actual voice section is small (i.e., when change of an ambient noise is small), it operates comparatively good. However, the noise source was a move object, the case where the distance to a noise source changes from the input edge of a sound signal, and the ambient noise were unsteady, when change was large, the presumed error with the noise spectrum actually superimposed on the presumed noise spectrum at voice became large, and there was a problem that a recognition performance deteriorated.

[0029] This invention aims at acquiring the voice recognition unit and method of being and cutting down the recognition performance degradation by change of the distance of the input edge of a sound signal, and a noise source for solving the above problems. Moreover, it aims at acquiring the voice recognition unit and method of cutting down the recognition performance degradation by change of an ambient noise.

[0030]

[Means for Solving the Problem] In the voice recognition unit which the voice recognition unit concerning this invention carries out the analysis of a spectrum of the noise superposition input sound signal including the non-voice section, and performs speech recognition processing in quest of a spectrum feature parameter A spectrum operation means to carry out the analysis of a spectrum of the noise superposition input sound signal, and to output noise superposition voice spectrum time series, An average spectrum operation means to presume the spectrum of superposition noise from the non-voice section in the noise superposition voice spectrum time series outputted from the above-mentioned spectrum operation means, and to output as a noise spectrum, The scale factor to the noise spectrum concerned at the time of subtracting the noise spectrum outputted from the above-mentioned average spectrum operation means from the noise superposition voice spectrum time series outputted from the above-mentioned spectrum operation means is changed. two or more kinds of normal-mode-rejection voice spectrum time series A normal-mode-rejection spectrum group operation means to output, and a feature-vector group operation means to change into two or more kinds of feature-vector time series two or more kinds of normal-mode-rejection voice spectrum time series outputted from the above-mentioned normal-mode-rejection spectrum group operation means, The collating model memory which comes to memorize the model showing changes of the kind of the noise-less voice pattern learned using the voice data uttered under environment without noise, and feature vector, To two or more kinds of normal-mode-rejection voice feature-vector time series outputted from the above-mentioned feature-vector group operation means in the 3-dimensional space which consists of time, a state, and three shafts of the kind of feature vector It is characterized by having a 3-dimensional collating means to perform collating with the model showing changes of the kind of the noise-less voice pattern memorized by the above-mentioned collating model memory and feature vector, and to output a recognition result.

[0031] Moreover, the noise spectrum outputted from the above-mentioned average spectrum operation means, And it has further the noise spectrum memory which memorizes two or more kinds of noise spectrum patterns beforehand learned using the clustering technique from a lot of noise data. Two or more kinds of scale factors to the above-mentioned noise vector from each noise superposition voice spectrum of the noise superposition voice spectrum time series to which the above-mentioned normal-mode-rejection spectrum operation means is outputted from the above-mentioned spectrum operation means, It is characterized by searching for two or more kinds of normal-mode-rejection voice spectrums combining two or more kinds of noise spectrum patterns memorized by the above-mentioned noise spectrum memory.

[0032] Moreover, the above-mentioned collating model memory is characterized by memorizing the model which does not add restrictions to changes of the kind of feature vector as a model showing changes of the kind of feature vector.

[0033] Moreover, the above-mentioned collating model memory is elgotic which can change in all kinds as a model which does not add restrictions to changes of the kind of feature vector. It is characterized by memorizing a hidden Markov model.

[0034] Moreover, the above-mentioned collating model memory is characterized by memorizing the model which added restrictions to changes of the kind of feature vector as a model showing changes of the kind of feature vector.

[0035] Moreover, the above-mentioned collating model memory is characterized by memorizing the hidden Markov model to which between the kinds of adjoining feature vector can change as a model which added restrictions to changes of the kind of feature vector.

[0036] Moreover, the speech recognition method concerning this invention is set to the speech recognition method of carrying out the analysis of a spectrum of the noise superposition input sound signal including the non-voice section, and performing speech recognition processing in quest of a spectrum feature parameter. The spectrum operation process of performing a analysis of a spectrum to noise superposition input voice, and obtaining noise superposition voice spectrum time series, The average spectrum operation process which presumes the spectrum of superposition noise from the non-voice section in the noise superposition voice spectrum time series obtained at the above-mentioned spectrum operation process, and is acquired as a noise spectrum, The normal-mode-rejection spectrum group operation process of changing the scale factor to the noise spectrum concerned at the time of subtracting the noise spectrum acquired from the noise superposition voice spectrum time series obtained at the above-mentioned spectrum operation process at the above-mentioned average spectrum operation process, and obtaining two or more kinds of normal-mode-rejection voice spectrum time series, The feature-vector group operation process of changing into two or more kinds of feature-vector time series two or more kinds of normal-mode-rejection voice spectrum time series obtained at the above-mentioned normal-mode-rejection spectrum group operation process, To two or more kinds of normal-mode-rejection voice feature-vector time series obtained at the above-mentioned feature-vector group operation process in the 3-dimensional space which consists of time, a state, and three shafts of the kind of feature vector It is characterized by having the 3-dimensional collating process of performing collating

with the model showing changes of the kind of the noise-less voice pattern learned using the voice data uttered under environment without noise, and feature vector, and obtaining the recognition result.

[0037] Moreover, the above-mentioned normal-mode-rejection spectrum operation process is characterized by searching for two or more kinds of normal-mode-rejection voice spectrums combining two or more kinds of scale factors to the above-mentioned noise vector, and two or more kinds of noise spectrum patterns beforehand learned using the clustering technique from a lot of noise data from each noise superposition voice spectrum of the noise superposition voice spectrum time series obtained at the above-mentioned spectrum operation process.

[0038] Moreover, the above-mentioned 3-dimensional collating process is characterized by using the model which does not add restrictions to changes of the kind of feature vector as a model showing changes of the kind of feature vector.

[0039] Moreover, the above-mentioned 3-dimensional collating process is elgotic which can change in all kinds as a model which does not add restrictions to changes of the kind of the above-mentioned feature vector. It is characterized by using a hidden Markov model.

[0040] Moreover, the above-mentioned 3-dimensional collating process is characterized by using the model which added restrictions to changes of the kind of feature vector as a model showing changes of the kind of feature vector.

[0041] Furthermore, the above-mentioned 3-dimensional collating process is characterized by using the hidden Markov model to which between the kinds of adjoining feature vector can change as a model which added restrictions to changes of the kind of feature vector.

[0042]

[Embodiments of the Invention] Form 1. drawing 1 of operation is the block diagram showing the composition for explaining the voice recognition unit and method concerning the form 1 of implementation of this invention. A spectrum operation means the same portion as the conventional example shown in drawing 8 shall attach and show the same sign in drawing 1 , and 101 performs a analysis of a spectrum to noise superposition voice input, and extract noise superposition voice spectrum time series, and 102 are average spectrum operation meanses average the spectrum of the non-voice section in the noise superposition voice spectrum time series outputted from the above-mentioned spectrum operation means 101, and output as a noise spectrum.

[0043] Moreover, as a new sign, 201 changes the scale factor to the noise spectrum at the time of subtracting the noise spectrum outputted from the above-mentioned average spectrum operation means 102 from the noise superposition voice spectrum time series outputted from the above-mentioned spectrum operation means 101, and subtracts a noise spectrum. A normal-mode-rejection spectrum group operation means to output two or more kinds of normal-mode-rejection spectrum time series, A feature-vector group operation means by which 202 changes two or more kinds of normal-mode-rejection spectrum time series into two or more kinds of feature-vector time series, 203 to two or more kinds of normal-mode-rejection voice feature-vector time series outputted from the above-mentioned feature-vector group operation means 202 in the 3-dimensional space which consists of time, a state, and three shafts of the kind of feature vector A 3-dimensional collating means to perform collating with the model showing changes of the kind of the noise-less voice pattern which the collating model memory 205 mentioned later memorizes, and feature vector, and to output a recognition result, 205 is collating model memory which comes to memorize the model showing changes of the kind of the noise-less voice pattern learned using the voice data generated under environment without noise, and feature vector.

[0044] Although the voice recognition unit concerning the form 1 of operation shown in this drawing 1 is constituted by the block diagram shown in drawing 1 mentioned above, it is equipped with the process shown below as a process which constitutes the corresponding speech recognition method.

a. The spectrum operation process of performing a analysis of a spectrum to noise superposition input voice, and obtaining noise superposition voice spectrum time series, b. The average spectrum operation process which presumes the spectrum of superposition noise from the non-voice section in the noise superposition voice spectrum time series obtained at the above-mentioned spectrum operation process, and is acquired as a noise spectrum, c. The scale factor to the noise spectrum concerned at the time of subtracting the noise spectrum acquired from the noise superposition voice spectrum time series obtained at the above-mentioned spectrum operation process at the above-mentioned average spectrum operation process is changed. two or more kinds of normal-mode-rejection voice spectrum time series The normal-mode-rejection spectrum group operation process to acquire, the feature-vector group operation process of changing into two or more kinds of feature-vector time series two or more kinds of normal-mode-rejection voice spectrum time series obtained at the d. above-mentioned normal-mode-rejection spectrum group operation process, e. to two or more kinds of normal-mode-rejection voice feature-vector time series obtained at the above-mentioned feature-vector group operation process in the 3-dimensional space which consists of time, a state, and three shafts of the kind of feature vector The 3-dimensional collating process of performing collating with the model showing changes of the kind of the noise-less voice pattern learned using the voice data uttered under environment without noise, and feature vector, and obtaining the recognition result.

[0045] Next, operation of the form 1 of operation concerning the above-mentioned composition is explained. Since operation of the spectrum operation means 101 and the average spectrum operation means 102 is the same as that of operation of the conventional example, it omits explanation here. With the normal-mode-rejection spectrum group operation means 201, using V kinds (two or more kinds) of sub TORAKUTO coefficient alpha (k) and (1 <=k<=V), a noise spectrum is subtracted and V kinds of normal-mode-rejection voice spectrum S (k) and (omega) are calculated from each noise superposition voice spectrum of the time series of a noise superposition voice spectrum. Here, the value of alpha (k) is set as 0.5 serration as follows.

[0046]
[Equation 10]

$$S^{(k)}(\omega) = \max\left\{X(\omega) - \alpha^{(k)}N(\omega), 0\right\}, \quad \alpha^{(k)} = 0.5k, \quad 1 \le k \le V \qquad (10)$$

[0047] Here, power [in / the frequency omega of the k-th kind of normal-mode-rejection voice spectrums / in S (k) and (omega)] and X (omega) express the power in the frequency omega of a noise superposition voice spectrum. Thus, V kinds of normal-mode-rejection voice spectrum time series S (1) and (omega), S (2) and (omega), ..., S (v) (omega) (However, S(k) (omega) = (S1 (k), (omega), S2 (k) and (omega), ..., ST (k), (omega)))
********.
[0048] V kinds of normal-mode-rejection voice spectrum time series S (1) which the normal-mode-rejection spectrum group operation means 201 outputs with the feature-vector group operation means 202, (omega), S (2), (omega), ..., V kinds of feature-vector time series Y that expresses the acoustical feature for S (v) and (omega) in speech recognition, such as an LPC cepstrum, like the conventional example (1), It changes into Y (2), ..., Y (v) (however, Y(k) =Y1(k), Y2 (k), ..., YT (k)).
[0049] With the 3-dimensional collating processing means 203, it collates to V kinds of feature-vector time series Y (1) which the feature-vector group operation means 202 outputs, Y (2), ..., Y (v) in the 3-dimensional space which consists of time, a state, and three shafts of the kind of feature vector, and the recognition candidate who gives a maximum likelihood is outputted as a recognition result.
[0050] The elgoticHMM model shown in drawing 2 expresses changes of the kind of feature vector. In drawing 2 , ckl is the transition probability to the kind l of the kind k of feature vector to feature vector, and it is connected with the Naru changes which do not output an observation event between each state. The elgoticHMM model is used for not attaching restrictions to changes of the kind of feature vector with the form 1 of this operation.
[0051] In order to find the optimal state and one sequence [become the likelihood maximum / of the combination of the kind of feature vector] (q, v) = (q1, v1), (q2, v2), ..., (qT, vT), the Viterbi search which consists of the following four steps and which was extended to three dimensions is performed.
[0052] STEP1 (initialization)
[0053]
[Equation 11]

$$\delta_1(i,k) = \pi_i \rho_k b_i(\mathbf{y}_1^{(k)}), \quad 1 \le i \le N, \quad 1 \le k \le V \qquad (11)$$

[0054]
[Equation 12]

$$\psi_t(i,k) = (0,0), \quad 1 \le i \le N, \quad 1 \le k \le V \qquad (12)$$

[0055] STEP2 (repeat)
[0056]
[Equation 13]

$$\delta_t(j,l) = \max_{1 \le i \le N, 1 \le k \le V} [\delta_{t-1}(i,k) a_{ij} c_{kl}] b_j(\mathbf{y}_t^{(l)}),$$

$$2 \le t \le T, \quad 1 \le j \le N, \quad 1 \le l \le V \qquad (13)$$

[0057]
[Equation 14]

$$\psi_t(j,l) = \arg\max_{1 \le i \le N, 1 \le k \le V} [\delta_{t-1}(i,k) a_{ij} c_{kl}],$$

$$2 \le t \le T, \quad 1 \le j \le N, \quad 1 \le l \le V \qquad (14)$$

[0058] STEP3 (end)
[0059]
[Equation 15]

$$P^* = \max_{1 \le i \le N, 1 \le k \le V} [\delta_T(i,k)] \qquad (15)$$

[0060]
[Equation 16]

$$(q_T^*, v_T^*) = \arg\max_{1 \le i \le N, 1 \le k \le V} [\delta_T(i,k)] \qquad (16)$$

[0061] STEP4 (backtracking)
[0062]
[Equation 17]

$$(q_t^*, v_t^*) = \psi_{t+1}(q_{t+1}^*, v_{t+1}^*), \quad t = T-1, T-2, \cdots, 1 \qquad (17)$$

[0063] Here, deltat (i, k) is a maximum likelihood in Time t, State i, and the kind k of feature vector on the path of one in the 3-dimensional space which consists of time, a state, and three shafts of the kind of feature vector, and is expressed with the following formulas.
[0064]
[Equation 18]

$$\delta_t(i,k) = \max_{q_1,q_2,\cdots q_{t-1},v_1,v_2,\cdots v_{t-1}} P[q_1 q_2 \cdots q_{t-1}, q_t = i, v_1 v_2 \cdots v_{t-1}, v_t = k, \mathbf{y}_1^{(1)}\mathbf{y}_2^{(1)}\cdots\mathbf{y}_t^{(1)}, \mathbf{y}_1^{(2)}\mathbf{y}_2^{(2)}\cdots\mathbf{y}_t^{(2)}, \cdots, \mathbf{y}_1^{(V)}\mathbf{y}_2^{(V)}\cdots\mathbf{y}_t^{(V)} \mid \lambda]$$

(18)

[0065] Formula (11) In - (17), psit (j, l) is two-dimensional array which memorizes the argument of the path which makes a formula (18) the maximum by each time t, each state j, and the kind l of feature vector. Moreover, as for the transition probability to the kind l of the kind k of feature vector to feature vector, and rhok, the kind of feature vector of the output probability of the feature vector [in / State i / in bi (yt (k))] yt (k) and ckl is the probability which is k in an initial state.
[0066] Drawing 3 is the state transition of the voice pattern for collating Left-to-right The situation of a 3-dimensional Viterbi search in case a HMM model expresses and an elgotic HMM model expresses changes of the kind of feature vector is expressed.
[0067] Moreover, drawing 4 is drawing which extracted the range of the time t-1 in drawing 3 - t. Maximum-likelihood deltat-1 [in / time t-1, State j, and the kind k of feature vector / in maximum-likelihood deltat (j, l) in Time t, State j, and the kind l of feature vector] (correcting (1 <=k<=V)) (j, k), Time t-1, a state j-1, the kind k of feature vector (calculating by choosing a path which becomes the likelihood maximum from maximum-likelihood deltat-1 (j-1, k) (correcting (1 <=k<=V)) which can be set is shown.)
[0068] Hereafter, the operation effect over the gestalt 1 of operation is described. In the conventional bottom voice recognition unit of noise, it assumed that the noise spectrum presumed from the non-voice section was uniformly overlapped on the whole tone voice section, and the value of the sub TORAKUTO coefficient alpha of ****** adjusted so that a recognition performance might become the maximum to evaluation data was used. However, since the power of the noise spectrum superimposed on voice in a certain time differs from the power of the noise spectrum at the time of noise presumption in changing the distance of a noise source and a voice input edge with time, a noise spectrum is lengthened too much, or it happens ** or that there is nothing too much, and an exact normal-mode-rejection voice spectrum cannot be searched for. As the result, a mismatch with a noise-less voice pattern occurs, and a recognition rate deteriorates.
[0069] Reference "the speech recognition under the unsteady noise noise by the parallel HMM method and the spectrum subtraction" (it ******) an electronic-intelligence communication society paper magazine (D-II), Vol.J-78-D-II, No.7, and pp.1021- in 1027 and 1995 Noise HMM is expressed by elgotic HMM and the recognition performance under unsteady noise environment is raised to the normal-mode-rejection voice feature vector after a spectrum subtraction by performing collating processing on the 3-dimensional space of the state of time and a voice model, and the state of a noise model. However, with the gestalt 1 of that there is no description about the value of a sub TORAKUTO coefficient, and this implementation, since changes of not a noise model but the kind of feature vector are modeled, both can tell the above-mentioned reference that it is another technology.
[0070] By the voice recognition unit and method concerning the form 1 of this operation, V kinds of feature-vector candidates who used and calculated V kinds of sub TORAKUTO coefficient alpha (k) at each time t of every exist. Since the kind k of feature vector in each time t is chosen so that a likelihood may serve as the maximum, even if it changes the distance of a noise source and a voice input edge, it can lengthen a noise spectrum too much, or it can prevent ** or there being nothing too much, and can suppress degradation of a recognition rate.
[0071] Moreover, although the elgotic HMM model which can change in all kinds is used as a model which expressed changes of the kind of feature vector with the voice recognition unit and method concerning the form 1 of this operation, without adding a limit to changes of the kind of feature vector By using the HMM model which shows between the kinds of feature vector which the value of sub TORAKUTO coefficient [at the time of a normal mode rejection] alpha (k) adjoins to drawing 5 whose changes were enabled as a model which added the limit to changes of the kind of feature vector It is possible to model a time change of superposition noise power appropriately.
[0072] Form 2. of operation, next drawing 6 are the block diagrams showing the composition for explaining the voice recognition unit and method concerning the form 2 of implementation of this invention. In drawing 6 , the same portion as the form 1 of operation shown in drawing 1 attaches and shows the same sign, and the explanation is omitted. As a new sign, 204

is noise spectrum memory which memorizes two or more kinds of noise spectrum patterns learned using the clustering technique from the noise spectrum outputted from the average spectrum operation means 102, and a lot of [beforehand] noise data. Two or more kinds of scale factors to the noise vector from each noise superposition voice spectrum of the noise superposition voice spectrum time series to which the normal-mode-rejection spectrum operation means 201 is outputted from the spectrum operation means 101, It is made combining two or more kinds of noise spectrum patterns memorized by the above-mentioned noise spectrum memory 204 as [search for / two or more kinds of normal-mode-rejection voice spectrums].

[0073] In addition, although the voice recognition unit concerning the gestalt 2 of operation is constituted by the block diagram shown in drawing 6 mentioned above As a process which constitutes the corresponding speech recognition method Two or more kinds of scale factors to the noise vector from each noise superposition voice spectrum of the noise superposition voice spectrum time series from which the normal-mode-rejection spectrum operation process concerning the gestalt 1 of operation mentioned above is acquired at a spectrum operation process, It is only differing combining two or more kinds of noise spectrum patterns beforehand learned using the clustering technique from a lot of noise data in that two or more kinds of normal-mode-rejection voice spectrums are searched for.
[0074] Next, operation of the form 2 of operation concerning the above-mentioned composition is explained. Since operation of the spectrum operation means 101 and the average spectrum operation means 102 is the same as that of operation of the conventional example, it omits explanation here. The noise spectrum which the average spectrum operation means 102 outputs by the noise spectrum memory 204, and V2 which were boiled and was beforehand learned using the clustering technique from a lot of noise data The representation noise spectrum pattern of a kind is memorized.
[0075] At the normal-mode-rejection spectrum group operation means 201, it is each noise superposition voice spectrum of the time series of a noise superposition voice spectrum to V1. The sub TORAKUTO coefficient alpha of a kind (k1), and (1 <=k1 <=V1), V2 the noise spectrum pattern Nk2 (omega) of a kind, and (1 <=k2 <=V2) -- combining -- a total of V -- = V1V2 Normal-mode-rejection voice spectrum [of a kind] S (k), (omega), and (1 <=k<=V) are calculated. Here, the value of alpha (k1) is set as 0.5 serration as follows.
[0076]
[Equation 19]

$$S^{(k)}(\omega) = \max\left\{X(\omega) - \alpha^{(k_1)} N_{k_2}(\omega), 0\right\},$$

$$\alpha^{(k_1)} = 0.5k_1, \quad 1 \le k_1 \le V_1, \quad 1 \le k_2 \le V_2, \quad 1 \le k \le V \qquad (19)$$

[0077] Here, power [in / the frequency omega of the k-th kind of normal-mode-rejection voice spectrums / in S (k) and (omega)], power / in / the frequency omega of a noise superposition voice spectrum / in X (omega)], and N (omega) express the power in the frequency omega of a presumed noise spectrum, respectively. Thus, V kinds of normal-mode-rejection voice spectrum time series S (1) and (omega), S (2) and (omega), ..., S (V) (omega) (however, it asks for S(k) (omega) = (S1 (k), (omega), S2 (k) and (omega), ..., ST (k), (omega)).)
[0078] Since operation of the feature-vector group operation means 202 and the 3-dimensional collating means 203 is the same as that of the gestalt 1 of operation, it omits explanation here.
[0079] Hereafter, the effect about the voice recognition unit and method concerning the gestalt 2 of operation is described. In the conventional bottom voice recognition unit of noise, the noise spectrum presumed from the non-voice section assumes that it superimposes on the whole tone voice section uniformly. however, unsteady noise environment, such as a run automatic in the car one, -- since the pattern of the noise spectrum superimposed on voice in a certain time differs from the pattern of the noise spectrum at the time of an average spectrum operation in changing as follows the pattern of the spectrum superimposed on voice with time, an exact normal-mode-rejection voice spectrum cannot be searched for A mismatch with a noise-less voice pattern occurs as the result, and a recognition rate deteriorates.
[0080] Moreover, by the voice recognition unit and method of a gestalt 1 of operation, although it can respond to change of spectrum power, since only a single noise spectrum pattern is used, it cannot respond about change of a spectrum pattern. With the voice recognition unit and method concerning the gestalt 2 of this operation, it is V1 in each time t of every. Sub TORAKUTO coefficients alpha (k1) and V2 of a kind V=V1V2 calculated using the noise spectrum pattern Nk2 (omega) of a kind The feature-vector candidate of a kind exists. The kind k of feature vector in each time t can suppress degradation of a recognition rate, even if it changes the noise spectrum pattern superimposed on the distance and voice of a noise source and a voice input edge, since it is chosen so that a likelihood may serve as the maximum.
[0081] Moreover, although the elgotic HMM model which can change in all kinds is used as a model which expressed changes of the kind of feature vector with the voice recognition unit and method concerning the gestalt 2 of this operation, without adding a limit to changes of the kind of feature vector As a model which added the limit to changes of the kind of feature vector, the noise spectrum pattern Nk2 (omega) at the time of a normal mode rejection is similar. Or it is possible to model appropriately a time change of a noise spectrum and a time change of superposition noise power by using the HMM model which shows between the kinds of feature vector which the value of sub TORAKUTO coefficient [at the time of a normal mode rejection] alpha (k) adjoins to drawing 7 whose changes were enabled.
[0082]

[Effect of the Invention] According to this invention, two or more kinds of feature-vector candidates who calculated using two or more kinds of sub TORAKUTO coefficients for every time exist, and as mentioned above, the kind of feature vector in each time Since it is chosen so that a likelihood may serve as the maximum, even if it changes the distance of a noise source and a voice input edge, lengthen a noise spectrum too much, or It can prevent that it is too ******, degradation of a recognition rate can be suppressed, and the recognition performance degradation by change of the distance of the input edge of a sound signal and a noise source can be cut down.

[0083] Moreover, even if it changes the noise spectrum pattern superimposed on voice, degradation of a recognition rate can be suppressed and the recognition performance degradation by change of an ambient noise can be cut down.

[0084] Moreover, degradation of a recognition rate can be suppressed by using the model which does not add a limit to changes of the kind of feature vector as a model showing changes of the kind of feature vector.

[0085] Moreover, degradation of a recognition rate can be suppressed by using the elgotic HMM model which can change in all kinds as a model which does not add a limit to changes of the kind of feature vector.

[0086] Moreover, a time change of superposition noise power can be appropriately modeled by using the model which added the limit to changes of the kind of feature vector as a model showing changes of the kind of feature vector.

[0087] Furthermore, a time change of superposition noise power can be appropriately modeled by using the HMM model which between the kinds of adjoining feature vector made changed as a model which added the limit to changes of the kind of feature vector.

[Translation done.]

CLAIMS

[Claim(s)]

[Claim 1] The voice recognition unit which carries out the analysis of a spectrum of the noise superposition input sound signal including the non-voice section characterized by providing the following, and performs speech recognition processing in quest of a spectrum feature parameter. A spectrum operation means to carry out the analysis of a spectrum of the noise superposition input sound signal, and to output noise superposition voice spectrum time series. An average spectrum operation means to presume the spectrum of superposition noise from the non-voice section in the noise superposition voice spectrum time series outputted from the above-mentioned spectrum operation means, and to output as a noise spectrum. A normal-mode-rejection spectrum group operation means to change the scale factor to the noise spectrum concerned at the time of subtracting the noise spectrum outputted from the above-mentioned average spectrum operation means from the noise superposition voice spectrum time series outputted from the above-mentioned spectrum operation means, and to output two or more kinds of normal-mode-rejection voice spectrum time series. A feature-vector group operation means to change into two or more kinds of feature-vector time series two or more kinds of normal-mode-rejection voice spectrum time series outputted from the above-mentioned normal-mode-rejection spectrum group operation means, The collating model memory which comes to memorize the model showing changes of the kind of the noise-less voice pattern learned using the voice data uttered under environment without noise, and feature vector, To two or more kinds of normal-mode-rejection voice feature-vector time series outputted from the above-mentioned feature-vector group operation means in the 3-dimensional space which consists of time, a state, and three shafts of the kind of feature vector A 3-dimensional collating means to perform collating with the model showing changes of the kind of the noise-less voice pattern memorized by the above-mentioned collating model memory and feature vector, and to output a recognition result.

[Claim 2] The noise spectrum outputted from the above-mentioned average spectrum operation means in a voice recognition unit according to claim 1, And it has further the noise spectrum memory which memorizes two or more kinds of noise spectrum patterns beforehand learned using the clustering technique from a lot of noise data. Two or more kinds of scale factors to the above-mentioned noise vector from each noise superposition voice spectrum of the noise superposition voice spectrum time series to which the above-mentioned normal-mode-rejection spectrum operation means is outputted from the above-mentioned spectrum operation means, The voice recognition unit characterized by searching for two or more kinds of normal-mode-rejection voice spectrums combining two or more kinds of noise spectrum patterns memorized by the above-mentioned noise spectrum memory.

[Claim 3] It is the voice recognition unit characterized by memorizing the model which does not add restrictions to changes of the kind of feature vector as a model with which the above-mentioned collating model memory expressed changes of the kind of feature vector in the voice recognition unit according to claim 1 or 2.

[Claim 4] It is elgotic which can change in all kinds as a model with which the above-mentioned collating model memory does not add restrictions to changes of the kind of feature vector in a voice recognition unit according to claim 3. Voice recognition unit characterized by memorizing a hidden Markov model.

[Claim 5] It is the voice recognition unit characterized by memorizing the model which added restrictions to changes of the kind of feature vector as a model with which the above-mentioned collating model memory expressed changes of the kind of feature vector in the voice recognition unit according to claim 1 or 2.

[Claim 6] It is the voice recognition unit characterized by memorizing the hidden Markov model to which between the kinds of adjoining feature vector can change as a model with which the above-mentioned collating model memory added restrictions to changes of the kind of feature vector in the voice recognition unit according to claim 5.

[Claim 7] The speech recognition method of carrying out the analysis of a spectrum of the noise superposition input sound signal including the non-voice section characterized by providing the following, and performing speech recognition processing in quest of a spectrum feature parameter. The spectrum operation process of performing a analysis of a spectrum to noise superposition input voice, and obtaining noise superposition voice spectrum time series. The average spectrum operation process which presumes the spectrum of superposition noise from the non-voice section in the noise superposition voice spectrum time series obtained at the above-mentioned spectrum operation process, and is acquired as a noise spectrum. The normal-mode-rejection spectrum group operation process of changing the scale factor to the noise spectrum concerned at the time of subtracting the noise spectrum acquired from the noise superposition voice spectrum time series obtained at the above-mentioned spectrum operation process at the above-mentioned average spectrum operation process, and obtaining two or more kinds of normal-mode-rejection voice spectrum time series. The feature-vector group operation process of changing

into two or more kinds of feature-vector time series two or more kinds of normal-mode-rejection voice spectrum time series obtained at the above-mentioned normal-mode-rejection spectrum group operation process, To two or more kinds of normal-mode-rejection voice feature-vector time series obtained at the above-mentioned feature-vector group operation process in the 3-dimensional space which consists of time, a state, and three shafts of the kind of feature vector The 3-dimensional collating process of performing collating with the model showing changes of the kind of the noise-less voice pattern learned using the voice data uttered under environment without noise, and feature vector, and obtaining the recognition result.

[Claim 8] It is the speech-recognition method characterized by to search for two or more kinds of normal-mode-rejection voice spectrums combining two or more kinds of scale factors to the above-mentioned noise vector, and two or more kinds of noise spectrum patterns beforehand learned using the clustering technique from a lot of noise data from each noise superposition voice spectrum of the noise superposition voice spectrum time series from which the above-mentioned normal-mode-rejection spectrum operation process is acquired at the above-mentioned spectrum operation process in the speech recognition method according to claim 7.

[Claim 9] It is the speech recognition method characterized by using the model which does not add restrictions to changes of the kind of feature vector as a model with which the above-mentioned 3-dimensional collating process expressed changes of the kind of feature vector in the speech recognition method according to claim 7 or 8.

[Claim 10] It is elgotic which can change in all kinds as a model with which the above-mentioned 3-dimensional collating process does not add restrictions to changes of the kind of the above-mentioned feature vector in the speech recognition method according to claim 9. The speech recognition method characterized by using a hidden Markov model.

[Claim 11] Claim 7 ** is the speech recognition method characterized by using the model which added restrictions to changes of the kind of feature vector as a model with which the above-mentioned 3-dimensional collating process expressed changes of the kind of feature vector in the speech recognition method given in 8.

[Claim 12] It is the speech recognition method characterized by using the hidden Markov model to which between the kinds of adjoining feature vector can change as a model with which the above-mentioned 3-dimensional collating process added restrictions to changes of the kind of feature vector in the speech recognition method according to claim 11.

[Translation done.]

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 04-318900
(43)Date of publication of application : 10.11.1992

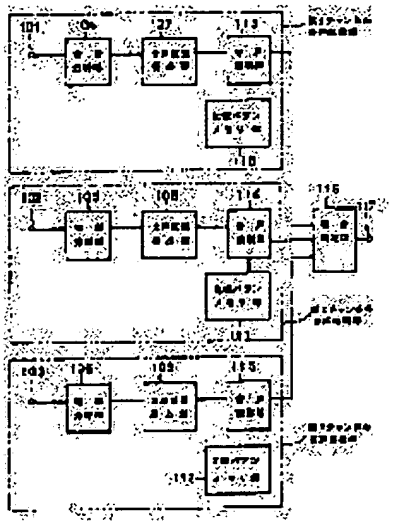
(51)Int.Cl. G10L 3/00
G10L 3/00
G10L 3/02

(21)Application number : 03-086645
(22)Date of filing : 18.04.1991

(71)Applicant : OKI ELECTRIC IND CO LTD
(72)Inventor : MIKI TAKASHI

(54) MULTIDIRECTIONAL SIMULTANEOUS SOUND COLLECTION TYPE VOICE RECOGNIZING METHOD

(57)Abstract:
PURPOSE: To obtain stable recognition performance even in environment wherein the distance and direction between the voicing windpipe and a microphone change and background noise environment changes.
CONSTITUTION: Voices which are collected through plural microphones at the same time are inputted from input terminals 101, 102, and 103 and passed through voice analysis parts 104, 105, and 106 and voice section detection part 107, 108, and 109, and comparison pattern memory parts 110, 111, and 112 are referred to, so that voice identification parts 113, 114, and 115 recognize them independently of one another. A total decision part 116 totally decides the results of the independent recognition and identification auxiliary information (identification accuracy, start and end times of voice, and signal-to-noise ratio) and outputs the final recognition result to an output terminal 117.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]